CSCI 620/Section 01/Mior, Introduction to Big Data, Fall 2201

Assignment 7 – Frequent itemset mining

Description

In this assignment, we are going to implement the Apriori algorithm using SQL with IMDB data. Your code will generate a new table for each level in the lattice. Here our items will be the set of actors and our “transactions” will be movies which actors have appeared in together. We will use a minimum support of 5.

Note: We will be ignoring the “pruning” step of the apriori-gen algorithm. This makes the code less efficient, but easier to implement, and we will have the same results.

Your tasks

1. Create a table Popular_Movie_Actors which is a subset of Movie_Actor containing only movies with type ‘movie’ and avgRating greater than 5. Provide your code. **(5 points)**

2. Provide SQL to create a table L1 which contains frequent itemsets of size one. Your table should have two columns: actor1 which contains the actor in the itemset, and count which contains the number of movies from Popular_Movie_Actors that this actor has appeared in. Make sure you only include itemsets which meet the minimum support. **(10 points)**

3. Provide SQL to create a table L2, which contains frequent itemsets of size two (actors who have appeared in the same movie together) with columns actor1, actor2, and count (the number of times these actors appeared in the same movie). This should be based only on your table L1 and Popular_Movie_Actors. Note that this can be written as a single SQL query. **(20 points)**

4. Provide SQL to create a table L3, which contains frequent itemsets of size three with columns actor1, actor2, actor3, and count. This should be based only on your table L2 and Popular_Movie_Actors. **(20 points)**

5. Write a program which generalizes your approach to Q3 and Q4 to generate tables for all levels of the lattice (L1, L2, …, Ln where n is the final level of the lattice). You should start with your code for Q2 which generates L1.

   You will need to programmatically generate and execute queries to create subsequent levels of the lattice (i.e. CREATE TABLE AS SELECT…). You should stop when you create an empty table at the final level of the lattice.

   (continued on next page)

Include the number of frequent itemsets in each level of the lattice in your report. For the last (non-empty) level of the lattice, include the names of the actors in each frequent itemset (by joining the table representing the final level of the lattice with the Members table). You can perform this final query manually. **(45 points)**